Bioinformatics Quick Start

Medha Bhagwat
National Center for Biotechnology Information
National Institutes of Health

Outline

1. What is a genome?
2. What is genomics?
3. What is bioinformatics?
4. Applications of genomics/bioinformatics
5. Future implications
6. A practical example
Proteins are Body's Worker Molecules

Hemoglobin carries oxygen to every part of the body

Ion channel proteins control brain signaling by allowing small molecules into and out of nerve cells

Enzymes in saliva, the stomach and the small intestine are proteins that help you digest food

Muscle proteins called actin and myosin enable all muscular movement

Antibodies are proteins that help defend your body against foreign invaders such as bacteria and viruses
DNA

Basic Unit (alphabet): Nucleotide (base)
Only 4: A, T, G, and C
Double-stranded

..AGCTGCATGCTAGCTGACGTCA....

||| ||| ||| ||| ||| ||| ||| ||| ||| ..TCGACGTACGATCGACTGCAGT....

“Words” (genes) to encode proteins, RNA etc.
Double helical

The double-helical structure of DNA
Protein

Alphabet: amino acids

There are 20 amino acids
Encoded by codons (triplets of nucleotides)

ATGTGCAGCCTAGCTGCGTC

Met—Cys—Ser—Leu—Ala—Ala—Val

Water channel protein
Genome (DNA)

Exact spelling of a word is necessary

<table>
<thead>
<tr>
<th>CAT</th>
<th>DAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAT</td>
<td>GAT</td>
</tr>
<tr>
<td>MAT</td>
<td>KAT</td>
</tr>
<tr>
<td>FAT</td>
<td>CBT</td>
</tr>
<tr>
<td>BAT</td>
<td>CCT</td>
</tr>
<tr>
<td>EAT</td>
<td>CAQ</td>
</tr>
<tr>
<td>HAT</td>
<td>CAC</td>
</tr>
</tbody>
</table>

Some changes in amino acids lead to diseases and some indicate normal differences among humans.

Genomics

Proteomics

http://genomics.energy.gov/
Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   - Photoreceptors

Additional Information

- [Genetics Home Reference](http://ghr.nlm.nih.gov/)
- [A Science Primer](http://www.ncbi.nlm.nih.gov/About/primer/)
- [Talking Glossary of Genetic Terms](http://www.genome.gov/10002096)
Bioinformatics

Variety of definitions
By Luscombe et al Method Inform Med 2001; 40:346-58

Bioinformatics is conceptualizing biology in terms of molecules
(in the sense of Physical chemistry)
and applying “informatics techniques”
(derived from disciplines such as applied
math, computer science and statistics)
to understand and organize the information associated
with these molecules, on a large scale.

Bioinformatics is a management information system for
molecular biology and has many practical applications.

Bioinformatics

I. Organize data in databases
   researchers can access current data
   submit new data

II. Develop tools and resources to analyze data

III. Interpret data in a biologically useful manner
    global analysis of data to uncover common
    principles that apply across many systems
National Center for Biotechnology Information


Created as a part of NLM in 1988
- To establish public databases
  U.S. National DNA Sequence Database
- To perform research in computational biology
- To develop software tools for sequence analysis
- To disseminate biomedical information

NCBI Databases

Primary Databases
- Genomes (DNA)
  - mRNA
    - Protein
  - Organisms

Derived Databases
- Expression
- Structures
- Small compounds
- Conserved Domains
- Publications
- Books
- Gene
- Homologene
- Unigene
- RefSeq
## NCBI Databases

<table>
<thead>
<tr>
<th>Primary</th>
<th>Derived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival/repository</td>
<td>Curated</td>
</tr>
<tr>
<td>Non-redundant</td>
<td>NCBI owner</td>
</tr>
<tr>
<td>Sequenced</td>
<td>Combined/edited</td>
</tr>
</tbody>
</table>

Ex: GenBank  
Ex: RefSeq

Genome Sequence Data and Analysis Tools at NCBI

Sizes of Different Genomes

Aloe vera 16.0 billion
Rabbit 3.5 billion
Human 3.2 billion
Laboratory mouse 2.6 billion
Fruit fly 137 million
Yeast 12.1 million
Bacterium (E. coli) 4.6 million
Human immunodeficiency virus 9700

Tools for Data Mining

<table>
<thead>
<tr>
<th>BLAST</th>
<th>OMIM</th>
<th>Entrez</th>
<th>Gene Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity Proteins</td>
<td>OMIM</td>
<td>Entrez</td>
<td>Gene Structure</td>
</tr>
<tr>
<td>Protein Sequence Analysis</td>
<td>OMIM</td>
<td>Entrez</td>
<td>Gene Structure</td>
</tr>
<tr>
<td>Nucleotide Sequence Analysis</td>
<td>OMIM</td>
<td>Entrez</td>
<td>Gene Structure</td>
</tr>
</tbody>
</table>

Genome Sequence Data and Analysis Tools at NCBI

**Tools**

- **Gene Expression**
  - GEO 
  - ArrayExpress 
  - SAGEmap 

- **Protein Expression**
  - Prosite 
  - UniProt

- **BLAST**
  - Basic Local Alignment Search Tool (BLAST)
  - TBLASTX
  - PSI-BLAST
  - BLAST2sequences

- **CD Search**
  - Search Conserved Domain Database with Reverse Position Specific BLAST

- **Gene Ontology (GO) Search**
  - Search for Gene Ontology terms in a database

- **SAGEmap**
  - SAGEmap is a tool for performing statistical tests designed for differential expression analysis of SAGE expression profiling data

- **SAGEmap)**
  - Comparison of SAGE libraries using a variety of scoring methods

- **The Cancer Genome Anatomy Project (CGAP)**
  - Aims to describe the molecular anatomy of cancer cells

- **ArrayCGH**
  - Digital Comparative Hybridization - online tool to compare expression profiles between selected SAGE libraries

- **ArrayCGH**
  - A new approach to the analysis of gene expression data

- **ArrayCGH**
  - A new approach to the analysis of gene expression data

- **ArrayCGH**
  - A new approach to the analysis of gene expression data

Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   How to access the genome data?
   How to access the analysis tools?
4. Applications of genomics/bioinformatics
   Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   Photoreceptors

Applications of Genomics and Proteomics

1. Understand basic biology
2. Diagnosis and treatment of diseases
3. Rationale for drug design
4. Protect plant life
5. Understand bacterial resistance
6. Solve environmental problems
7. Develop new energy sources
8. Improve industrial processes
9. Study evolutionary changes among organisms
Human Genome Sequenced

Celera

The Human Genome Project

June 23, 2000

The Human Genome

23 pairs of chromosomes
3.2 billion base pairs
Estimated number of genes about 30,000
Only 2% of the human genome “codes”
Average gene size 4000 base pairs
Largest gene dystrophin 2.4 million base pairs
More than 50% in repeat elements or so called “junk DNA”
Analysis of the Human Genome

The DNA sequence of any two people is 99.9 percent identical.

Sites in the DNA sequence where individuals differ at a single DNA base are called single nucleotide polymorphisms (SNPs).

The SNPs may greatly affect an individual's disease risk.

Sickle Cell Anemia

- Sickled red blood cells
- Mutation in the HBB gene that codes for hemoglobin
- one nucleotide change in the 7th codon GAG to GTG
- changing glutamic acid to valine
- interaction between valine and the complementary regions on adjacent molecules results in the formation of polymers that aggregate and distort the shape of the red blood cells

3-D structure of hemoglobin
3-D structure of mutant hemoglobin
Human Genome Sequenced

What are the genetic changes that make us human?

Only 1.2% difference between Human and chimpanzee:
Every 100th nucleotide different
Could affect thousands of genes
Many are probably the consequence of 6 million years
of genetic drift, with little effect on body or behavior
Other small changes--perhaps in regulatory, noncoding
sequences--may have dramatic consequences

What are the genetic changes that make us human?

Study clinical mutations in the genes that impair key traits
and trace the genes' evolution.
For example, FOXP2, the first ever gene associated with
human speech.
FOXP2 knock-out mice are squeak less.

A complete understanding of uniquely human traits will,
however, include more than DNA such as nurture and
nature.
Understand Bacterial Resistance

Fluoroquinolone antibiotics kill Tuberculosis bacteria by binding to DNA-DNA gyrase complex. Tuberculosis bacterium encodes a novel protein mfpA resembling DNA. mfpA competes with DNA for binding to Fluoroquinolone antibiotics thus making bacteria resistant to the antibiotic.

**DNA**

mfpA protein

*Science (2005) 308, 1393*

---

*Photo: © UC Berkeley Electron Microscope Lab (GNN)*

---

*Methanocaldococcus jannaschii*

Methane-producing thermophilic archeon

Produces methane, an important energy source. Encodes enzymes that withstand high temperatures and pressures possibly useful for industrial processes.

---

*Photo: © UC Berkeley Electron Microscope Lab (GNN)*
**Thalassiosira pseudonana**

Ocean diatom, a major participant in biological pumping of carbon to ocean depths

has potential for mitigating global climate change

Photo: courtesy of DOE-Genomes to Life

---

**Deinococcus radiodurans**

Survives extremely high levels of radiation

has high potential for radioactive waste cleanup

Photo: DOE Joint Genome Institute
What’s Next?????

1. HapMap: Genetic variation mapping project
   Discovery of genes related to diseases
   Gene Testing
   Gene Therapy
2. Pharmacogenomics: Pharmacology and genomics
   Custom effective drugs based on genetic profile
   Reduce adverse reactions
3. ENCODE: Encyclopedia of functional elements
   Study expression of genes

Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   How to access the genome data?
   How to access the analysis tools?
4. Applications of genomics/bioinformatics
   Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   Photoreceptors
More Information

Volume 291, Issue 5507, Pages 1145-1434

Nature Supplement on “Human Genome”
http://www.nature.com/nature/supplements/collections/humangenome/index.html

Human Genome Project Information
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

NHGRI Fact sheets
http://www.genome.gov/10000202

Obtain more information about opsin genes and proteins

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhodopsin</td>
<td>Blue, Rod photoreceptor, rhodopsin receptor</td>
</tr>
<tr>
<td>Hemoglobin subunits alpha</td>
<td>Red, Globin chain, oxygen transporter</td>
</tr>
<tr>
<td>Chlorophyll a</td>
<td>Green, Photosynthesis pigment</td>
</tr>
</tbody>
</table>

**Entrez Gene**

- **Rhodopsin**
  - **Function**: Blue, Rod photoreceptor, rhodopsin receptor
  - **Chromosome**: 1p12, 1p21
  - **Gene ID**: 7066

- **Hemoglobin subunits alpha**
  - **Function**: Red, Globin chain, oxygen transporter
  - **Chromosome**: 16q22
  - **Gene ID**: 3277

- **Chlorophyll a**
  - **Function**: Green, Photosynthesis pigment
  - **Chromosome**: 16p13
  - **Gene ID**: 3276

**GenBank Accessions**

- **Rhodopsin**: NC_000001.11 (MIR_20622), NC_000005.10 (MIR_60665)
- **Hemoglobin subunits alpha**: NC_000001.11 (MIR_1990)
- **Chlorophyll a**: NC_000001.11 (MIR_1990)
Information about Rhodopsin Gene RHO

Human Vision
Absorption of light by photoreceptor cells in eye

<table>
<thead>
<tr>
<th>Rods</th>
<th>Cones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noncolor vision in dim light</td>
<td>color vision in bright light</td>
</tr>
<tr>
<td>pigment opsin (chromophore retinal)</td>
<td>rhodopsin blue green red</td>
</tr>
<tr>
<td>Absorbs at 500 nm</td>
<td>426nm 530nm 560nm</td>
</tr>
<tr>
<td>gene RHO</td>
<td>OPN1SW OPN1MW OPN1LW</td>
</tr>
<tr>
<td>Chromosome 3</td>
<td>7 X X</td>
</tr>
<tr>
<td>Protein length 348</td>
<td>348 362 362</td>
</tr>
<tr>
<td>Rest 40% identity</td>
<td>95% identity</td>
</tr>
</tbody>
</table>
- Only 15 amino acids different
- 3 residues determine the wavelength of absorption
  At 180 (serine/alanine), 277 (tyrosine/phenyl alanine), 285 (threonine/alanine)
- Hydroxyl containing amino acids in the red pigment interact with the photo excited state of retinal and lower its energy, leading to a shift toward the lower-energy (red) region of the spectrum
- Mutations in these amino acids lead to color “blindness”

Red and Green Genes Susceptible to Unequal Homologous Recombination

- High level of identity between them
- Positioned on chromosome X adjacent to each other
- Leading to different number of individual genes or hybrid genes and thus color “blindness”
- Trouble distinguishing red and green color
- Approximately 5% of males have only the red gene
Humans have three cone pigments red, green and blue.

Common Lineage Tree Generated from the Taxonomy Browser
Mice are not sensitive to light as far toward the infrared region and they do not discriminate colors well.

Birds have highly acute color perception.

Birds, for example, chickens have 4 cone pigments red, green, blue similar to humans and an additional one, violet.
Fish, for example, Zebrafish have multiple copies of 4 pigments.

Fishes have special receptor requirement because of the variation in the amount of light in water.
Blast Distance Tree of Animal Proteins Similar to Human Rhodopsin (contd)

Tax BLAST Report of the Previous BLAST Search
Green and red photoreceptors are products of a recent evolutionary event.

The green and red pigments appear to have diverged in the primate lineage approximately 35 million years ago.

Mammals, such as dogs and mice, that diverged from primates earlier have only two cone photoreceptors, blue and green, an event believed to have resulted form the nocturnal life.

In contrast, birds such as chickens have a total of six pigments: rhodopsin, four cone pigments, and a pineal visual pigment called *pinopsin*. Birds have highly acute color perception.

Aquatic environment offers a single system to study evolution of color vision because of the variations in underwater light.

Review articles:
Bowmaker and Hunt Current Biology vol16, R484
Hunt et al. CMLS, Cell.Mol.Lifr Sci. 58, 1583
Obtain More Information about Opsin Genes and Proteins

Information about Rhodopsin Gene RHO

Summary

Pigmentosa is an inherited progressive disease which is a major cause of blindness in western communities. It can be inherited as an autosomal dominant, autosomal recessive, or X-linked recessive disorder. In the autosomal dominant form, which comprises about 25% of total cases, approximately 30% of families have mutations in the gene encoding the rod photoreceptor-specific protein rhodopsin. This is the transmembrane protein which, when photoexcited, initiates the visual transduction cascade. Defects in this gene are also one of the causes of congenital stationary night blindness.
Automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes

<table>
<thead>
<tr>
<th>Genes identified as putative homologs of one another during the construction of HomoloGene</th>
<th>Proteins used in sequence comparisons and their conserved domain architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.247565: Expression restricted to eye</td>
<td>UniGene: Organized View of the Transcriptome</td>
</tr>
<tr>
<td>Restricted Expression (contributing more than half of the EST frequency)</td>
<td></td>
</tr>
<tr>
<td>Hs.247565: Expression restricted to eye</td>
<td></td>
</tr>
</tbody>
</table>

**Gene Details:**
- **H. sapiens Rho modopsin (opsin 2, rod pigment)**, human, NP_000501.1, 340 aa
- **C. tropiplus Rho modopsin (opsin 2, rod pigment)**, chimps, XP_617460.2, 340 aa
- **C. lupus Rho modopsin (opsin 2, rod pigment)**, dogs, XP_859208.1, 358 aa
- **M. musculus Rho modopsin**, mouse, NP_662628.1, 340 aa
- **P. troglodytes Rho modopsin**, rat, NP_354976.1, 340 aa
- **G. gallus Rho modopsin (opsin 2, rod pigment)**, chicken, NP_990221.1, 355 aa
- **C. roeni Rho modopsin**, zebrafish, NP_671158.1, 354 aa

**UniGene Cluster:**
- Cluster of transcript sequences that appear to come from the same gene/expressed pseudogene
Information about Rhodopsin Gene RHO

Human Vision

Opsins are 7 transmembrane helix receptors (7TM family)

Chromophore 11 cis-retinal covalently binds to lysine (296) to form positively charged Schiff base

A positive Schiff base is compensated by glutamate(113)

On absorption of light isomerizes to 11 trans-retinal

Leads to cascade of events that cause hyperpolarization of the membrane and neuronal signaling
Blink: Precalculated top 200 protein BLAST hits

<table>
<thead>
<tr>
<th>Score</th>
<th>Accession</th>
<th>GI</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Query: C01249491 [Blast All]  
Name: C01249491  
Accession: C01249491  
GI: 96402938  
Species: Homo sapiens  
Features: 1. Proteins (62)  

240 hits to 1 unique species:
Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   - Photoreceptors
Bioinformatics

I. Organize data in databases
   researchers can access current data
   submit new data
II. Develop tools and resources to analyze data
III. Interpret data in a biologically useful manner
     global analysis of data to uncover common
     principles that apply across many systems

Vision for Bioinformatics

Databases
- Entrez global search
- Genomes
- Books
- Gene
- Nucleotide
- RefSeq
- Protein
- Taxonomy
- Homologene
- UniGene
- AceView
- dbSNP
- OMIM
- Structure

Tools
- BLAST
- Blastp
- Blast2 sequences
- Distance Tree
- Tax Blast
- MapViewer
- UniGene DDD
- Cn3D
- Taxonomy Common Tree
- Blink
- Related Structures

Interpretation
- Photoreceptors
- Cones and rods
- Sequence similarity
- Phylogeny
- Homology
- Expression
- Structure
- Function
Questions about NCBI Resources?

E-mail
info@ncbi.nlm.nih.gov